Environ Ecol Stat (2007) 14:27–40
DOI 10.1007/s10651-006-0004-x

ORIGINAL ARTICLE

Isomorphic chain graphs for modeling spatial dependence in ecological data

Alix I. Gitelman · Alan Herlihy

Received: 15 October 2004 / Revised: 11 March 2005 /
Published online: 20 January 2007

Abstract Graphical models (alternatively, Bayesian belief networks, path analysis models) are increasingly used for modeling complex ecological systems (e.g., Lee, In: Ferson S, Burgman M(eds) Quantative methods for conservation biology. Springer, Berlin Heilin Heideslperk New York, pp.127–147, 2000; Borsuk et al., J Water Res Plann Manage 129:271–282, 2003). Their implementation in this context leverages their utility in modeling interrelationships in multivariate systems, and in a Bayesian implementation, their intuitive appeal of yielding easily interpretable posterior probability estimates. However, methods for incorporating correlational structure to account for observations collected through time and/or space—features of most ecological data—have not been widely studied; Haas et al. (AI Appl 8:15–27, 1994) is one exception. In this paper, an “isomorphic” chain graph (ICG) model is introduced to account for correlation between samples by linking site-specific Bayes network models. Several results show that the ICG preserves many of the Markov properties (conditional and marginal dependencies) of the site-specific models. The ICG model is compared with a model that does not account for spatial correlation. Data from several stream networks in the Willamette River valley, Oregon (USA) are used. Significant correlation between sites within the same stream network is shown with an ICG model.

Keywords Bayesian belief network · Graphical model · Spatial correlation

A. I. Gitelman(✉)
Statistics Department, Oregon State University, Corvallis, OR, USA
e-mail:gitelman@stat.orst.edu

A. Herlihy
Fisheries and Wildlife Department, Oregon State University, Corvallis, OR, USA

1 Introduction

Graphical models (e.g., Bayesian belief networks, path analysis models) are increasingly used for modeling complex ecological systems (e.g., Lee 2000; Borsuk et al. 2003). Their implementation in this context leverages their utility in modeling directional (or influential) relationships in multivariate systems. For example, we might connect land-use characteristics to indicators of macro-invertebrate health in a stream, where the connection is influenced by water temperature, riparian condition and stream chemistry. Of course, we must temper any causal language with assumptions of no confounding factors and/or with scientific arguments regarding the ecological process. Figure 1 shows a hypothetical graphical model for the situation just described. Each node represents a random variable, and the directed edges between nodes indicate influences (in the absence of unmeasured confounding factors) of variables on each other. A Bayesian implementation—Bayes networks—has the added intuitive appeal of providing posterior probability estimates for the parameters of each node distribution and the parameters associated with each directed edge.

To estimate the parameters of a model like the one shown in Fig. 1, we typically assume multivariate observations that are independent and identically distributed. Independence between observations for ecological processes is particularly suspect, as these data are often collected in spatial and/or temporal proximity. Haas et al. (1994) incorporate serial correlation into a Bayes net model for Aspen stand growth by connecting nodes from within-year Bayes nets to subsequent within-year Bayes nets in a “feed-forward” fashion, resulting in a 12-year model with close to 500 nodal distributions.

As an alternative, we introduce isomorphic chain graphs (ICG) in which observations at different locations are assumed to have identical graphical structures, but observations that are close together in space are connected with a chain (or undirected) link between corresponding nodes. Figure 2 shows an isomorphic chain graph model for two observations at neighboring sites. The structure of the within-site models is the same as that of Fig. 1; we simply connect the two observations with a chain link (an undirected edge) at the health metric node. The ICG allows us to leverage similarities in underlying ecological processes across sites for estimating parameters of the graphical structure, while also accounting for correlation through space. We show that this ICG connection preserves many of the conditional and marginal independence properties from the unconnected (or independent) case, thereby preserving the influences between univariate components within each site.

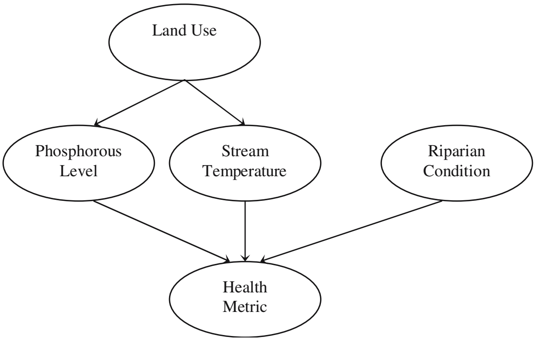

Fig. 1 Hypothetical graphical model showing relationships between variables in an ecological system

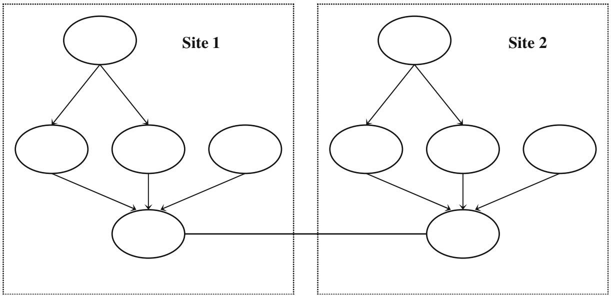

Fig. 2 Hypothetical isomorphic chain graph structure for two neighboring sites in an ecological system

Next, we describe data sampled from several stream networks in the Willamette Valley Ecoregion, Oregon. Using this example, we briefly catalog some graph theory terminology and notation in Sect. 3 (a more extensive catalog is provided in Appendix A). We formally define the ICG model in Sect. 4 and present the main results regarding conditional and marginal independencies. To illustrate the ICG model, in Sect. 5, we parameterize spatial correlation between observations at neighboring sites in the Willamette Valley data using a spatial autoregressive model (Ord, 1975). We fit several models to the Willamette Valley data and compare the results to a multiple linear regression model and a non-correlated graphical model (a Bayes network model). We conclude with several remarks on further extensions to the ICG model that may be effective in addressing additional spatial dependencies across complex multivariate observations.

2 Willamette river basin data

We use data collected from wadeable streams in the Willamette Valley as part of Environmental protection Agencies (EPAs) Environmental Monitoring and Assessment Program (EMAP) and Ag-Riparian Project. Data were collected in summer 1996 and 1997, as detailed in Van Sickle et al. (2004). In summer 1999, as part of the second phase of the Ag-Riparian project, five of the sites were selected for an intensive longitudinal upstream sampling program with from five to seven new sample sites in each of the upstream networks at intervals of 1 – 3 km. Figure 3 shows the original sample locations in the Willamette Valley, as well as the site distribution in each of the five intensively sampled stream networks. It is these fine sampling interval stream networks that motivate the extension from Bayes network models to ICG models. The original stream sampling sites are far enough apart in space so that any spatial correlation between sites is likely to be negligible; it is the within-network dependence that we model here.

The Willamette Valley dataset is rich with information on stream dynamics and riparian condition, as well as on the biological condition of stream denizens; Table 1 provides a partial list of these measurements. To illustrate the isomorphic chain graph and its properties, we focus on a ratio of observed-to-expected macro-invertebrate taxa richness for disperser taxa as a stream health index. This measure indicates a detrimental impact on the macro-invertebrate community if the ratio is

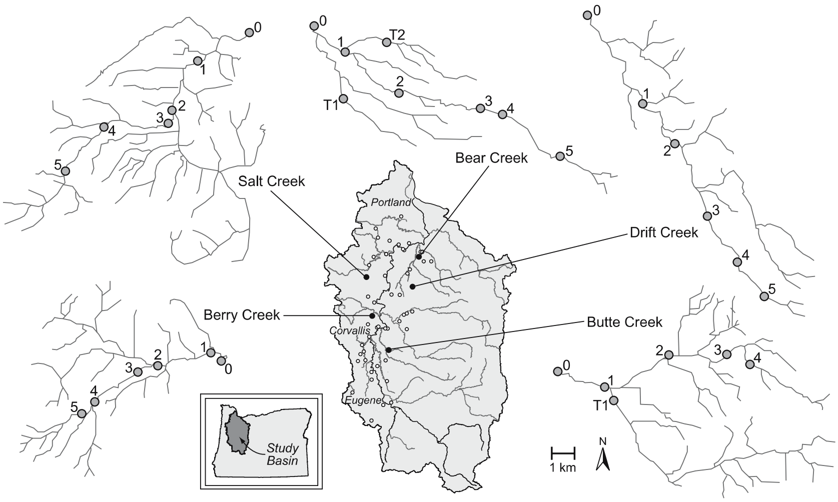

Fig. 3 Map of Willamette Valley showing sampling locations

Table 1 Partial list of variables in the Willamette Valley Data

Variable	Description
Agricultural	Percent of agricultural land cover within 150 m wide upstream network buffer
Developed	Percent of developed land cover within 150 m wide upstream network buffer
Channel simplification	Essentially, a coefficient of variation for mean stream depth
Phosphorous	Total phosphorous content of the sample
Temperature	Temperature of the water sample
Riparian condition	An index of riparian habitat quality
Watershed area	km^2; upstream drainage area
Elevation	Meters above sea level
Stream power	calculated as (stream slope × (watershed area)$^{1/4}$
Longitude	
Latitude	
DOE	Marcoinvertebrate dispersers observed-to-expected ratio

substantially less than one. Predictive modeling is a widely used approach that evaluates the biotic integrity of a sampled site by comparing its observed biota to the biota to be expected if the site were in reference condition and minimally altered by human activities (Wright et al. 1993). Clarke et al. (1996) give statistical details of the predictive modeling approach.

Our "dispersers" observed-to-expected ratio (DOE) is essentially the same as the Willamette invertebrate observed/expected index (WINOE) metric used in the assessment of Van Sickle et al. (2004), except that it uses only taxa that are considered to be dispersers—those taxa whose members emerge from the stream as adults and have the potential to fly some distance away to reproduce. Strongly dispersive taxa can move on the order of tens of meters up to a few kilometers. Downstream drift is also a major dispersal mechanism for all taxa in these streams. For comparison, we use variables identified by Van Sickle et al. (2004) for modeling the observed-to-expected index of macro-invertebrate health: percent agricultural land, percent developed land, longitude, and stream power.

3 Graphical model terminology and notation

Pictorially, a graph consists of nodes and edges. Directed edges are noted with arrows (as in $v \longrightarrow w$) and undirected edges by lines (as in $v - w$). Mathematically, a graph $\mathbf{G}$ is a pair of sets, $(\mathbf{V}, \mathbf{E})$, where $\mathbf{V} = \{v_1, v_2, \ldots, v_k\}$ denotes a finite set of vertices (nodes or univariate variables) and

$$\mathbf{E} = \{(v, w) : v, w, \in V \text{ and there is a directed edge from } v \text{ to } w\},$$

denotes a set of edges. If both of (v, w) and (w, v) are in $\mathbf{E}$, then there is an undirected edge between v and w.

A path is an ordered sequence of vertices connected by directed and/or undirected edges. If a path consists only of directed (undirected) edges it is called directed (undirected). In particular, a graph with only directed edges and no cycles (a cycle is a path leading to and from the same node) is called an acyclic directed graph (ADG). A chain graph (CG) has a combination of directed and undirected edges, but must have no directed or semi-directed cycles (a cycle in which at least one of the edges is directed). For an ADG, $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, and a subset $\mathbf{A}$ of $\mathbf{V}$, the parents of $\mathbf{A}$, denoted $\mathbf{Pa}(\mathbf{A})$, comprise the set of all nodes $v \in \mathbf{V}$, such that $(v, a) \in \mathbf{E}$ for some $a \in \mathbf{A}$.

The Markov factorization for ADG (Pearl 2000, p. 16) allows for expressing the joint probability distribution of nodes in an ADG as the product of each node, conditional only on its parents. That is, the joint probability distribution of the vertices can be written:

$$f(v_1, \ldots, v_k) = \prod_{j=1}^{k} f(v_j | \mathbf{Pa}(\mathbf{v_j})).$$

For our purposes, two ADG are said to be identically distributed if their vertex and edge sets correspond (this is stronger than saying that their probability distributions can be factored in the same manner).

Pearl (1988) introduced d-separation as a criterion for identifying conditional and marginal independence (equivalently, Markov factorizations) in ADG models, and Lauritzen et al. (1990) developed an algorithm for showing d-separation. Andersson et al. (2001) followed with AMP-separation, the parallel criterion and method for CG models (AMP stands for "alternative Markov property"). Definitions for d-separation and AMP-separation are provided in Appendix A, but their benefits, due to Verma and Pearl (1992) for ADG and to Andersson et al. (2001) for CG, are in providing criteria for identifying conditional and marginal independencies in ADG and CG models, respectively. We rely on AMP-separation for our main results in the next section.

4 Isomorphic chain graphs

To estimate the parameters of an ADG model such as the one in Fig. 1, we assume that the observations (samples) used to estimate both node and edge parameters are independent and identically distributed. The presumption is that the same graphical structure exists at all sites independently of other sites. In what follows, we relax this assumption, allowing for associational dependence across samples in what we term an "isomorphic" chain graph (ICG). In words, an ICG is a chain graph constructed by connecting identical ADG with undirected edges between corresponding univariate components or nodes; an example is shown in Fig. 2. In this way, the general structure of the ADG is preserved across sites, but we allow a spatial correlation between corresponding univariate components at those sites. A formal definition follows.

Let $\mathbf{G_1} = (\mathbf{V_1}, \mathbf{E_1})$; $\mathbf{G_2} = (\mathbf{V_2}, \mathbf{E_2}); \ldots ; \mathbf{G_n} = (\mathbf{V_n}, \mathbf{E_n})$ be identically distributed ADG, where $\mathbf{V_i} = \{v_{i1}, \ldots, v_{ik}\}$ for $i = 1, \ldots, n$ are ordered k-tuples. Then

$$\mathbf{G} = \left(\bigcup_{i=1,\ldots,n} \mathbf{V_i}, \bigcup_{i=1,\ldots,n} \mathbf{E_i} \cup \mathbf{E}_j^* \right),$$

where for some $j \in \{1, \ldots, k\}$,

$$\mathbf{E}_j^* = \{(v_{ij}, w_{i'j}), (w_{i'j}, v_{ij}) : v_{ij} \in \mathbf{V_i}, w_{i'j} \in \mathbf{V_{i'}} \text{ for } i, i' \in \mathbf{I}_j\}$$

is an ICG with corresponding index set, $\mathbf{I}_j$, which identifies the nodes with chain link connections. The nodes in $\mathbf{E}_j^*$ are called the isomorphic nodes.

In the hypothetical ICG model of Fig. 2 the isomorphic connection is made at the bottom of the two ADG components. Certainly, non-isomorphic chain link connections across within-site ADG might also be reasonable in the ecological modeling context. For instance, it might be reasonable to think that stream power at site s_1, upstream from site s_2, might be correlated with macro-invertebrate health at s_2. In this case, however, because there is a direction to streamflow, the feed-forward Bayes network model of Haas et al. (1994) seems more appropriate.

We now present two results regarding the conditional and marginal independencies associated with ICG models. Since an ICG model is constructed from identical ADG models, it seems appealing that some of the properties of the individual ADG models carry over to the full ICG. This is, in fact, the substance of our first result; the marginal and conditional independencies inherent in the individual ADGs also hold in the ICG. In our second result, we show that some nodes in individual ADG components are marginally independent of the corresponding nodes in other ADG components making up the ICG. In a related result, we establish conditional independencies across ACG components in the ICG model when we condition on the isomorphic nodes.

Together, our results allow a convenient factorization of the probability distribution described by the ICG model. We describe this factorization following the results themselves. To simplify the presentation of these results, we state them in terms of only two ADG components, although they are easily generalized to an arbitrary number of these components. Proofs (and heuristics for extension to arbitrary numbers of components) are provided in Appendix B. For both results, we rely on the following ICG construction. Let $\mathbf{G}_1 = (\mathbf{V}_1, \mathbf{E}_1)$ and $\mathbf{G}_2 = (\mathbf{V}_2, \mathbf{E}_2)$ be two identically distributed ADG. Let $\mathbf{G} = \left(\mathbf{V}_1 \cup \mathbf{V}_2, \mathbf{E}_1 \cup \mathbf{E}_2 \cup \mathbf{E}_j^*\right)$, where $\mathbf{E}_j^* = \{(v_{1j}, v_{2j}), (v_{2j}, v_{1j}) : v_{1j} \in \mathbf{V}_1 \text{ and } v_{2j} \in \mathbf{V}_2\}$, denote the ICG constructed from $\mathbf{G}_1$ and $\mathbf{G}_2$.

Result 1 For pairwise disjoint, nonempty subjects of $\mathbf{V}_1$ (equivalently, $\mathbf{V}_2$), $\mathbf{A}, \mathbf{B}, \mathbf{C}$,

$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C} \text{ in } \mathbf{G}_1(\mathbf{G}_2) \Longleftrightarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C} \text{ in } \mathbf{G}.$$

This result asserts that conditional independencies that hold in the marginal ADG also hold in the ICG.

For Result 2, we define an ancestral set as in Andersson et al. (2001). Consider the graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, and take $\mathbf{A} \subseteq \mathbf{V}$. A node $v \in \mathbf{V}$ is an ancestor of a node $a \in \mathbf{A}$ if there is a directed path from v to a in $\mathbf{G}$. The ancestral set of $\mathbf{A}$ is

$$\mathbf{An}(\mathbf{v}) \overset{\text{def}}{=} \mathbf{A} \cup \{v \in \mathbf{V} : v \text{ is an ancestor of } a \text{ for some } a \in \mathbf{A}\}.$$

That is, $\mathbf{An}(\mathbf{A})$ is that subset of $\mathbf{V}$ that contains $\mathbf{A}$ and all of its ancestors. Also, for notational convenience, we define the $\cap^*$ operator as follows. For a subset $\mathbf{A}$ of $\mathbf{V}_1 \cup \mathbf{V}_2$,

$$\left\{a_{1j}, a_{2j}\right\} \in \mathbf{A} \cap^* \mathbf{E}_j^*,$$

if $\left\{a_{1j}, a_{2j}\right\} \in \mathbf{A}$ and $(a_{1j}, a_{2j}) \in \mathbf{E}_j^*$.

Result 2

(a) For $\mathbf{A}_1 \subseteq \mathbf{V}_1$ and $\mathbf{A}_2 \subseteq \mathbf{V}_2$, if $\mathbf{An}(\mathbf{A_1} \cup \mathbf{A_2}) \cap^* \mathbf{E}_j^* = \emptyset$ then $\mathbf{A}_1 \perp\!\!\!\perp \mathbf{A}_2$ in $\mathbf{G}$.

(b) For $\mathbf{A}_1 \subseteq \mathbf{V}_1$ and $\mathbf{A}_2 \subseteq \mathbf{V}_2$ such that $(\mathbf{A}_1 \cup \mathbf{A}_2) \cap^* \mathbf{E}_j^* = \emptyset$ but $\mathbf{An}(\mathbf{A_1} \cup \mathbf{A_2}) \cap^* \mathbf{E}_j^* = \left\{a_{1j}, a_{2j}\right\}$:
  (1) $\mathbf{A}_1 \backslash \{a_{1k}\} \perp\!\!\!\perp \mathbf{A}_2 \backslash \{a_{2k}\} \mid \{a_{1k}, a_{2k}\}$.
  (2) $\mathbf{A}_1 \backslash \{a_{1k}\} \perp\!\!\!\perp a_{2k} \mid a_{1k}$.
  (3) $\mathbf{A}_2 \backslash \{a_{2k}\} \perp\!\!\!\perp a_{1k} \mid a_{2k}$.

Result 2(a) indicates that marginal independence across ADG components used to construct an ICG is maintained for those nodes in the ICG that lie above the isomorphic nodes. Result 2(b) indicates in part that corresponding univariate components below the isomorphic nodes in an ICG are conditionally independent, given the isomorphic nodes.

The benefit of Results 1 and 2 comes in allowing a convenient factorization of the joint probability distribution. From identical ADG $\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_n$, suppose we construct an ICG with isomorphic nodes $\mathbf{E}_m^*$. Then for $\mathbf{V}_1 = \{v_{11}, \ldots, v_{1k}\}, \ldots, \mathbf{V}_n = \{v_{n1}, \ldots, v_{nk}\}$, the joint probability distribution of $\mathbf{V}_1 \cup \cdots \cup \mathbf{V}_n$ is

$$f(\mathbf{V}_1, \ldots, \mathbf{V}_n)|\boldsymbol{\phi}) = \left[\prod_{i=1}^{n} \left\{\prod_{j=1}^{m-1} f(v_{ij}|\mathbf{Pa}(\mathbf{v_{ij}}))\right\} \left\{\prod_{j=m+1}^{k} f(v_{ij}|\mathbf{Pa}(\mathbf{v_{ij}}))\right\}\right]$$
$$\times f(v_{im}, \ldots, v_{nm}|\mathbf{Pa}(\{\mathbf{v_{1m}}, \ldots, \mathbf{v_{nm}}\})).$$

In the final term of this expression, we model the joint distribution of the isomorphic nodes conditional on their parents using a multivariate distribution. An implementation of this model is given next.

5 Bayes network and ICG parameterizations

To compare results with Van Sickle et al. (2004), we used the four explanatory variables they found to be most important for modeling the observed-to-expected ratio of macro-invertebrate species: longitude (L), stream power (P), percent agricultural development (A) and percent developed land (D); the latter two are measured within a 150-m buffer of the stream. We used H to denote the DOE index of macro-invertebrate health. The correlation between the WINOE ratio in Van Sickle et al. (2004) and our dispersers health metric, DOE, is quite high ($\hat{\rho} = 0.93, n = 76$). We used $n = 76$, having removed four samples with unusually large streamflow. It is also worth noting that while we would not think of longitude as "causing" a lowered DOE, there might be a relationship in which sites that are "poorer" for macro-invertebrates are more likely to be at the eastern end of the Willamette Valley.

We fit the multiple linear regression model as in Van Sickle et al. (2004), but using a Bayes Network approach. That is, we assigned probability distributions to all explanatory variables, excluding longitude (we assume longitude to be fixed for all models). This approach facilitates comparison with the Bayes network and ICG models, which also assign probability distributions to all nodes. For the ADG version of the multiple linear regression model, we took each explanatory variable to have a directed arrow into the response, with no arrows between explanatory variables.

Exploratory analysis indicates that both stream power (transformed by the quartic root following Van Sickle et al. (2004) and agricultural land cover follow Normal distributions. Imposing a Normality assumption on the percent developed land cover is more difficult—the distribution is highly skewed. We therefore modeled this variable using an exponential probability distribution. Our Bayes network multiple linear regression model (MLR) is

$$A_i \overset{iid}{\sim} N(\mu_a, \sigma_a^2), \qquad D_i \overset{iid}{\sim} Exp(\theta), \qquad P_i \overset{iid}{\sim} N(\mu_p, \sigma_p^2), \qquad H_i \overset{ind}{\sim} N(\gamma_i, \sigma^2), \tag{1}$$

where

$$\gamma_i = \beta_0 + \beta_1 L_i + \beta_2 A_i + \beta_3 U_i + \beta_4 P_i \tag{2}$$

for $i = 1, 2, \ldots n$.

For the Bayes network model (BN), where we model correlations between explanatory variables, we use Markov factorizations to write the distribution of all variables conditional on its parents. But first, we must decide on a structure for the BN. To do this, we use Tetrad 4.3 (Spirtes et al. 1993; http://www.phil.cmu.edu/tetrad, 2005), a graphical modeling freeware package. Tetrad assumes Normal distributions for all nodes, which is not our situation, though the resulting model is reasonable ecologically.

The graphical structure suggested by Tetrad is shown in Fig. 4. Interestingly, longitude is not pulled into the model at all, and there is no arrow from percent agriculture to DOE, whereas both percent developed and stream power have arrows into percent agricultural development and DOE. This makes some sense ecologically, in that stream power is a major driver of stream substrate size, which is in turn a major controlling factor for stream macro-invertebrates. Developed land in the network buffer may be related to agriculture, because as the landscape is urbanized (developed) in the Willamette Valley, it is usually by conversion of agricultural land. The major axis of the Willamette Valley is oriented north-south, with sites at the eastern and western edges of the Valley being steeper and less desirable for agriculture. This may help explain the absence of arrows from the longitude node.

To parameterize this model structure, we take

$$D_i \overset{iid}{\sim} Exp(\theta), \qquad P_i \overset{iid}{\sim} N(\mu_p, \sigma_p^2), \qquad A_i \overset{ind}{\sim} N(\delta_i, \sigma_a^2), \qquad H_i \overset{ind}{\sim} N(\gamma_i, \sigma^2), \tag{3}$$

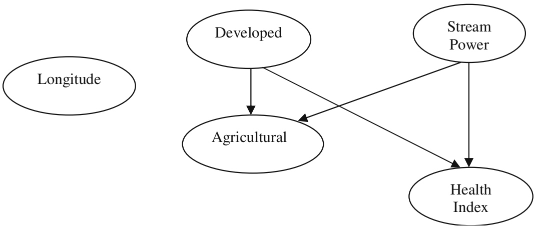

Fig. 4 Bayesian network suggested by Tetrad (2003)

where

$$\delta_i = \phi_0 + \phi_1 P_i + \phi_2 D_i, \tag{4}$$

$$\gamma_i = \beta_0 + \beta_1 P_i + \beta_2 D_i. \tag{5}$$

We now turn to the ICG models in which we parameterize the isomorphic nodes using a spatial autoregressive model. For sample location, s_i, for $i = 1, \ldots, n$, let N_i denote the neighborhood set for location s_i, in which $s_j \in N_i$ if s_j is in the same stream as s_i. For $s_j \in N_i$, we write $s_j \sim s_i$. We define an $n \times n$ weight matrix $\mathbf{W}$, in which each entry, w_{ij}, is defined as follows:

$$w_{ij} = \begin{cases} a_{ij} & : \quad s_i \sim s_j, \\ 0 & : \quad \text{otherwise}, \end{cases}$$

where

$$a_{ij} = \frac{\frac{1}{||s_i - s_j||}}{\sum_{s_k \in N_i} \frac{1}{||s_i - s_k||}},$$

using the Euclidean distance metric, $|| \cdot ||$. In this way, observations are defined as neighbors only if they are in the same intensively sampled stream network.

Following Congdon (2003, p. 253), to add an autoregressive component to the health metric node (this is ICG_1) we pre-multiply both the health metric and stream power and percent urban land cover by $\mathbf{W}$, and modify the model component of (5) to be

$$\gamma_i = \beta_0 + \beta_1 P_i + \beta_2 D_i + \rho W H_i - \rho\beta_1 W P_i - \rho\beta_2 W D_i. \tag{6}$$

Similarly, to add an autoregressive component to the agriculture node (this is ICG_2), we pre-multiply agriculture, stream power and urban by $\mathbf{W}$ and replace (5) with

$$\delta_i = \phi_0 + \phi_1 P_i + \phi_2 D_i + \rho W A_i - \rho\phi_1 W P_i - \rho\phi_2 W D_i. \tag{7}$$

5.1 Prior distributions

For the regression parameters $\boldsymbol{\beta}$and $\boldsymbol{\phi}$in each of these models, we use non-informative Normal priors. Similarly, for each variance term, σ^2, we use non-informative inverse-gamma priors. The more difficult parameterization involves the spatial parameter, ρ, in the ICG models. We restrict attention to $\rho \in (0, 1)$, as positive spatial correlation is more realistic within streams, and following Congdon (2003, Example 7.1), we use a uniform prior on the interval zero to one (a beta prior with parameters $\alpha = \beta = 1$). To compare the Bayes network and ICG models, we use the Bayesian information criterion (BIC; Schwarz, 1978).

6 Results for the Willamette Valley data

All models are fit using WinBUGS 1.4 (Spielgelhalter et al. 2003). The weighting **W** is fairly sparse in our case, as only five streams were intensively sampled. Therefore, we also examined (assuming isotropy) a correlogram of the residuals, which indicates positive spatial correlation $\hat{\rho} = 0.73$) at distances less than 4 km. This certainly includes distances between many of the samples taken from within one of the five intensively sampled streams. After a burn-in period, we assess convergence in WinBUGS by starting independent chains at different values and assessing the “scale reduction” (Gelman et al. 2004, p. 297) for each component of the chain.

In Table 2, we show results from fitting each of the four models: the MLR model, the BN model assuming independence between samples, and the two ICG models—ICG_1, with isomorphic nodes DOE, and ICG_2, with isomorphic nodes corresponding to the percent agriculture nodes. Notice the BN and ICG_2 match almost exactly. The estimated correlation corresponding to the percent agriculture nodes in ICG_2 is essentially zero, and so the model is virtually identical to BN. A significant, positive correlation is estimated under ICG_1, indicating the dependence of samples within the intensively sampled streams.

By fitting the MLR model as a BN model, we are able to compare it with the BN and ICG models using a likelihood-based criterion. Table 3 shows this comparison of the four models using BIC. The ICG model with the spatial autoregressive component on the health index is best by a small margin, followed by the BN model, and then the remaining two models. It does appear that the penalty for including an extra parameter for spatial dependence in the DOE nodes is offset by the gain in the likelihood component of the BIC.

Table 2 Results—posterior means (standard deviations)—from the multiple regression model (*MLR*), the Bayes network model assuming independence between samples (BN), and the two isomorphic chain graph models

Parameter	MLR	BN	ICG_1	ICG_2
μ_p	1.14 (0.03)	1.14 (0.03)	1.14 (0.04)	1.14 (0.04)
σ_p 0.30 (0.02)	0.30 (0.03)	0.30 (0.03)	0.30 (0.03)	
μ_a	42.7 (2.8)	NA	NA	NA
σ_a	24.3 (2.0)	20.6 (1.7)	20.6 (1.8)	20.8 (1.8)
ϕ_0	NA	87.58 (9.6)	87.63 (9.6)	85.74 (10.0)
ϕ_1	NA	−0.46 (0.11)	−0.46 (0.11)	−0.46 (0.12)
ϕ_2	NA	−35.04 (8.0)	−35.06 (8.0)	−34.12 (8.3)
β_0	31.58 (7.9)	0.14 (0.07)	0.13 (0.06)	0.14 (0.07)
β_1	−0.25 (0.06)	NA	NA	NA
β_2	−0.0023 (0.0007)	NA	NA	NA
β_3	−0.0032 (0.0008)	−0.0019 (0.0008)	−0.0014 (0.0008)	−0.0019 (0.0008)
β_4	0.30 (0.05)	0.34 (0.07)	0.31 (0.05)	0.34 (0.06)
σ	0.14 (0.01)	0.14 (0.01)	0.13 (0.01)	0.14 (0.01)
ρ	NA	NA	0.42 (0.16)	0.02 (0.02)
θ	0.09 (0.01)	0.09 (0.01)	0.09 (0.01)	0.09 (0.01)

ICG_1 denotes the model with isomorphic nodes corresponding to DOE and ICG_2 denote the model with isomorphic nodes corresponding to percent agriculture

Table 3 Comparison of the multiple linear regression model (*MLR*), the Bayes network model assuming independent samples (*BN*), and the two ICG models—ICG_1 has DOE isomorphic nodes and ICG_2 has percent agriculture isomorphic nodes

Model	BIC
MLR	1184
BN	1179
ICG_1	1174
ICG_2	1184

7 Discussion

The isomorphic chain graph model represents a useful and practical extension BN models that assume statistical independence between observations used to estimate parameters of the model. Specifically, two key conditional independence results are inherited from the BN paradigm, whereby the ICG model also affords a convenient factorization of the joint probability distribution depicted by the graphical model. In the examples described here, the substantive conclusions do not change with a change in the nodes modeled isomorphically; however, the overall variability explained by the different models does reflect their differences.

One possible extension to the ICG model incorporates several isomorphic nodes. Referring back to Figs. 1 and 2, we might consider chain graph connections at both the land use node and the health metric node. Connections between this approach and co-kriging might warrant further investigation, although the computational burden might be a constraining factor. Further, placing the isomorphic nodes at the "top" of the ICG, for example, at a land-use node, seems similar in spirit to a hierarchical structuring of nodes across sites. This connection is the topic of ongoing research.

Acknowledgements The authors are supported by Science To Achieve Results (STAR) Program Cooperative Agreement #CR-829095 from the US Environmental Protection Agency (EPA). We thank Steve Jensen for his assistance with some preliminary data analysis. Thanks also to Danny Lee and an anonymous reviewer for their comments and suggestions for improving the manuscript. We are also indebted to Scott Urquart for his support.

Appendix A

This notation and terminology is taken from Andersson et al. (2001).

Definition 1 (*Separation*) For a UDG, $\mathbf{GU} \equiv (\mathbf{V_u}, \mathbf{E_u})$, and for non-empty, pairwise disjoint subsets $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ of $\mathbf{V_u}$, $\mathbf{A}$ and $\mathbf{B}$ are said to be *separated* by $\mathbf{C}$ in $\mathbf{GU}$ if all paths in $\mathbf{GU}$ between $\mathbf{A}$ and $\mathbf{B}$ pass through $\mathbf{C}$. $\mathbf{A}$ and $\mathbf{B}$ are separated in $\mathbf{GU}$ if there are no paths between them in $\mathbf{GU}$.

Definition 2 (*Subgraph*) A graph $\mathbf{G}' \equiv (\mathbf{V}', \mathbf{E}')$ is called a *subgraph* of a graph $\mathbf{G} \equiv (\mathbf{V}, \mathbf{E})$ if $\mathbf{V}' \subseteq \mathbf{V}$ and $\mathbf{E}' \subseteq \mathbf{E}$. A subset $\mathbf{A}$ of $\mathbf{V}$ *induces* a subgraph, denoted $\mathbf{G_A} \equiv (\mathbf{A}, \mathbf{E_A})$, where $\mathbf{E_A}$ is that subset of edges in $\mathbf{E}$ with both endpoints in $\mathbf{A}$.

Definition 3 (*Coherent set*) For two nodes $v, w \in \mathbf{V}$, v is *coherent* to w if there is an undirected path in $\mathbf{G}$ between v and w. The *coherent set* of a set $\mathbf{A} \subseteq \mathbf{V}$ is

$$\mathbf{Co}(\mathbf{A}) \overset{\text{def}}{=} \mathbf{A} \cup \{v \in \mathbf{V} : v \text{ is coherent to } a \text{ for some } a \in \mathbf{A}\}.$$

$\mathbf{Co}(\mathbf{A})$ is that subset of $\mathbf{V}$ that contains $\mathbf{A}$ and all nodes coherent to $\mathbf{A}$.

Definition 4 (*Undirected subgraph*) For a graph $\mathbf{G} \equiv (\mathbf{V}, \mathbf{E})$, the *undirected subgraph* is

$$\mathbf{G}^* \equiv (\mathbf{V}, \mathbf{E}^*),$$

where

$$\mathbf{E}^* = \{(v, w) : (v, w) \in \mathbf{E} \text{ and } (w, v) \in \mathbf{E}\}.$$

Definition 5 (*Extended subgraph*) For an ADG, $\mathbf{G} \equiv (\mathbf{V}, \mathbf{E})$ with pairwise disjoint subsets $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ of $\mathbf{V}$, the *extended subgraph* is

$$\mathbf{G}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}] \overset{\text{def}}{=} \mathbf{G}_{\mathbf{An}(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C})}.$$

For a CG $\mathbf{G} \equiv (\mathbf{V}, \mathbf{E})$ with pairwise disjoint subsets $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ of $\mathbf{V}$, the *extended subgraph* is

$$\mathbf{G}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}] \overset{\text{def}}{=} \mathbf{G}_{\mathbf{An}(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C})} \cup \mathbf{G}^*_{\mathbf{Co}(\mathbf{An}(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}))},$$

where for $\mathbf{G1} \equiv (\mathbf{V_1}, \mathbf{E_1})$ and $\mathbf{G2} \equiv (\mathbf{V_2}, \mathbf{E_2})$, $\mathbf{G1} \cup \mathbf{G2} \equiv (\mathbf{V_1} \cup \mathbf{V_2}, \mathbf{E_1} \cup \mathbf{E_2})$.

Definition 6 (*Augmented graph*) For an ADG or CG $\mathbf{G} \equiv (\mathbf{V}, \mathbf{E})$, the *augmented graph*, denoted $\mathbf{G^a}$, is an UDG constructed according to:

(1) identify all flags and bi-flags (see Fig. 5);
(2) moralize all flags and bi-flags (see Fig. 6);
(3) replace all directed edges with undirected edges (see Fig. 6).

Figure 5 shows *flags* in (a) – (c). These are ordered triples of nodes with the given configurations. If $\mathbf{G}$ is an ADG, then the only kind of flag is that in Fig. 5a. Figure 5(d) encodes four different configurations of *bi-flags*, in that ? can be replaced by a directed edge in either direction, an undirected edge or no edge. *Moralizing* flags and bi-flags is accomplished by adding undirected edges where no edges exist (in the three- or four-node system). Figure 6 shows the moralized versions of the flags and bi-flags of Fig. 5.

Definition 7 (*d-separation*) For an ADG $\mathbf{G} \equiv (\mathbf{V}, \mathbf{E})$ and pairwise disjoint subsets of $\mathbf{V}$, $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, if $\mathbf{A}$ and $\mathbf{B}$ are separated by $\mathbf{C}$ in $\mathbf{G}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}]^{\mathbf{a}}$, then $\mathbf{A}$ and $\mathbf{B}$ are d-separated by $\mathbf{C}$ in $\mathbf{G}$. If $\mathbf{A}$ and $\mathbf{B}$ are separated in $\mathbf{G}[\mathbf{A} \cup \mathbf{B}]^{\mathbf{a}}$, then $\mathbf{A}$ and $\mathbf{B}$ are d-separated in $\mathbf{G}$. Pearl's result then implies the conditional (marginal) independencies.

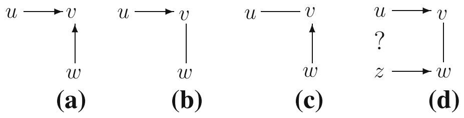

Fig. 5 **(a)**, **(b)**, **(c)** flags; **(d)** a bi-flag

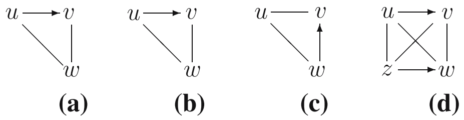

Fig. 6 Augmented versions of the flags in Figures 5a,b,c and the bi-flags in Figure 5d

Definition 8 (*AMP-separation*) Suppose $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are pairwise disjoint subsets of $\mathbf{V}$ in a CG, $\mathbf{G} \equiv (\mathbf{V}, \mathbf{E})$. Then if $\mathbf{C}$ separates $\mathbf{A}$ and $\mathbf{B}$ in $\mathbf{G}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}]^{\mathbf{a}}$, this is called AMP-separation and $\mathbf{A}$ and $\mathbf{B}$ are conditionally independent given $\mathbf{C}$, denoted $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \big| \mathbf{C}$. In the special case that $\mathbf{C} = \emptyset$, then $\mathbf{A}$ and $\mathbf{B}$ are marginally independent if there are no paths connecting them in $\mathbf{G}[\mathbf{A} \cup \mathbf{B}]^{\mathbf{a}}$.

Appendix B

Proof of Result 1 First suppose that for subsets $\mathbf{A}, \mathbf{B}, \mathbf{C}$ of $\mathbf{V}_1$, $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \big| \mathbf{C}$ in $\mathbf{G}$. Then $\mathbf{A}$ and $\mathbf{B}$ are separated by $\mathbf{C}$ in $\mathbf{G}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}]^{\mathbf{a}}$. Since $\mathbf{G_1}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}]$ is an ADG subgraph of $\mathbf{G}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}]$, it follows that $\mathbf{G_1}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}]^{\mathbf{a}}$ is a CG subgraph of $\mathbf{G}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}]^{a}$. As any subgraph has at most as many edges as the original graph, it must be that $\mathbf{C}$ separates $\mathbf{A}$ and $\mathbf{B}$ in $\mathbf{G_1}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}]^{\mathbf{a}}$, whereby $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \big| \mathbf{C}$ in $\mathbf{G_1}$. Next suppose that $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \big| \mathbf{C}$ in $\mathbf{G_1}$. Because $\mathbf{G_1}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}] = \mathbf{G}_{\mathbf{An}(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C})}$, write

$$\mathbf{G}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}] = \mathbf{G_1}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}] \cup \mathbf{G}^*_{\mathbf{Co}(\mathbf{An}(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}))}.$$

If $\mathbf{An}(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}) \cap^* \mathbf{E}^*_j = \emptyset$, then $\mathbf{G}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}] = \mathbf{G_1}[\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}]$ and the result holds. On the other hand, if $\mathbf{An}(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}) \cap^* \mathbf{E}^*_j \neq \emptyset$, then any flag introduced by including $\mathbf{G}^*_{\mathbf{Co}(\mathbf{An}(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}))}$ must of the kind in Fig. 5b or c. Therefore, no "moralizing" of the graph will result in a connection across subgraphs. Therefore, Result 1 holds.

Proof of Result 2 For (a), as there are no nodes $(a_{1j}, a_{2j}) \in \mathbf{An}(\mathbf{A_1} \cup \mathbf{A_2})$, such that $(a_{1j}, a_{2j}) \in \mathbf{E}^*_{\mathbf{j}}$, $\mathbf{G}[\mathbf{A_1} \cup \mathbf{A_2}] = \mathbf{G}_{\mathbf{An}(\mathbf{A_1} \cup \mathbf{A_2})} = \mathbf{G_1}[\mathbf{An}(\mathbf{A_1})] \cup \mathbf{G_2}[\mathbf{An}(\mathbf{A_2})]$. There are no flags or bi-flags connecting nodes in $\mathbf{A_1}$ and $\mathbf{A_2}$ in $\mathbf{G}[\mathbf{An}(\mathbf{A_1} \cup \mathbf{A_2})]$, so that augmenting the graph results only in changing directed edges to undirected edges, whereby, $\mathbf{A_1}$ and $\mathbf{A_2}$ are AMP-separated in $\mathbf{G}$. For (b), it is useful to notice that if $v \in \mathbf{An}(\mathbf{A_1} \cup \mathbf{A_2}) \cap \mathbf{A_1} \backslash \{a_1\}$, then necessarily, $(a_1, v) \in \mathbf{E_1}$. Similarly, if $w \in \mathbf{An}(\mathbf{A_1} \cup \mathbf{A_2}) \cap \mathbf{A_2} \backslash \{a_2\}$, then $(a_2, w) \in \mathbf{E_2}$. Hence, the only path between v and w in $\mathbf{G}[\mathbf{A_1} \cup \mathbf{A_2}]$ takes the form, $v \longleftarrow a_1 - a_2 \longrightarrow w$, which is *not* a bi-flag. Therefore, in the augmented graph $\mathbf{G}[\mathbf{A_1} \cup \mathbf{A_2}]^{\mathbf{a}}$, the only path between v and w must pass through both a_1 and a_2, whereby a_1 and a_2 AMP-separate v and w in $\mathbf{G}$. This completes the proof.

Heuristic for extending Results 1 and 2 to n iid ADG: this is essentially a proof by induction. What remains is to evaluate flags and bi-flags introduced by adding an additional ADG component to an existing ICG. But these will affect only the results in the way described in the proof of Result 2—that is, both Results still hold.

References

Andersson SA, Madigan D, Perlman MD (2001) An alternative Markov property for chain graphs. Scand J Stat 28:33–86

Borsuk ME, Stow CA, Reckhow KH (2003) Integrated approach to total maximum daily load development for Neuse River Estuary using Bayesian probability network model (Neu-BERN). J Water Res Plann Manage 129:271–282

Clarke RT, Furse MT, Wright JF, Moss D (1996) Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna. J Appl Stat 23:311–332

Congdon P (2003) Applied Bayesian modelling. Wiley, West Sussex, UK

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis 2nd ed. Chapman and Hall, London

Haas TC, Mowrer HT, Sheppard WD (1994) Modeling aspen stand growth with a temporal Bayes network. AI Appli 8:15–27

Lauritzen SL, Dawid AP, Larsen BN, Leimer H-G (1990) Independence properties of directed Markov fields. Networks 20:491–505

Lee D (2000) Assessing land-use impacts on bull trout using Bayesian belief networks. In: Ferson S, Burgman M (eds) Quantitative methods for conservation biology. Springer, Berlin Heildelbeg New York, pp 127–147

Ord K (1975) Estimation of methods for models of spatial interaction. J Am Stat Assoc 70:120–126

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo

Pearl J (2000) Causality: models, reasoning and inference, University of Cambridge, Cambridge, UK

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 5:461–464

Spielgelhalter D, Thomas A, Best N, Lunn D (2003) WinBUGS user manual. http://www.mrc-bsu.cam.ac.uk/bugs, last accessed April 2005

Spirtes P, Glymour C, Scheines R (1993) Causation, prediction, and search. Springer, Berlin Heidelberg New York; Tetrad 4.3, Department of Philosophy, Carnegie Mellon University, http://www.phil.cmu.edu/tetrad, last accessed April 2005

Van Sickle J, Baker J, Herlihy A, Bayley P, Gregory S, Haggerty P, Ashkenas L, Li J (2004) Projecting the biological condition of streams under alternative scenarios of human land use. Ecol Appl 14:368–380

Verma T, Pearl J (1992) An algorithm for deciding if a set of observed independencies has a causal explanation. In: Dubois D, Wellman M, D'Ambrosio B, Smets P, (eds) Proceedings of the eighth conference on uncertainty in artificial intelligence. Morgan Kaufman, San Francisco, pp 323–330

Wright JF, Furse MT, Armitage PD (1993) RIVPACS: a technique for evaluating the biological water quality of rivers in the UK. Eur Water Pollut Control 3:15–25

Biographical sketches

Alix Gitelman is Associate Professor of Statistics at Oregon State University. Her main research interests are graphical models and Bayesian statistics. Her work includes applications in environmental and educational statistics.

Alan Herlihy is Senior Research Associate Professor of Fisheries and Wildlife at Oregon State University. He is trained in environmental science and experienced in analyzing environmental water chemistry. He has been associated with the EPA's Enivronmental Monitoring and Assessment Program (EMAP) since its inception, and has had a major role in organizing EMAP's data bases.